Análisis Inteligente de Datos: Segundo Parcial

Claudio Sebastián Castillo

04 de mayo de 2022

Intro ANOVA

Datos

```
## spec_tbl_df [25 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ VitaminaB: num [1:25] 7.9 6.2 6.6 8.6 8.9 10.1 9.6 5.7 7.5 9.8 ...
##  $ Marca    : num [1:25] 1 1 1 1 1 1 1 2 2 2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   VitaminaB = col_double(),
##   ..   Marca = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
## # A tibble: 0 x 0
```

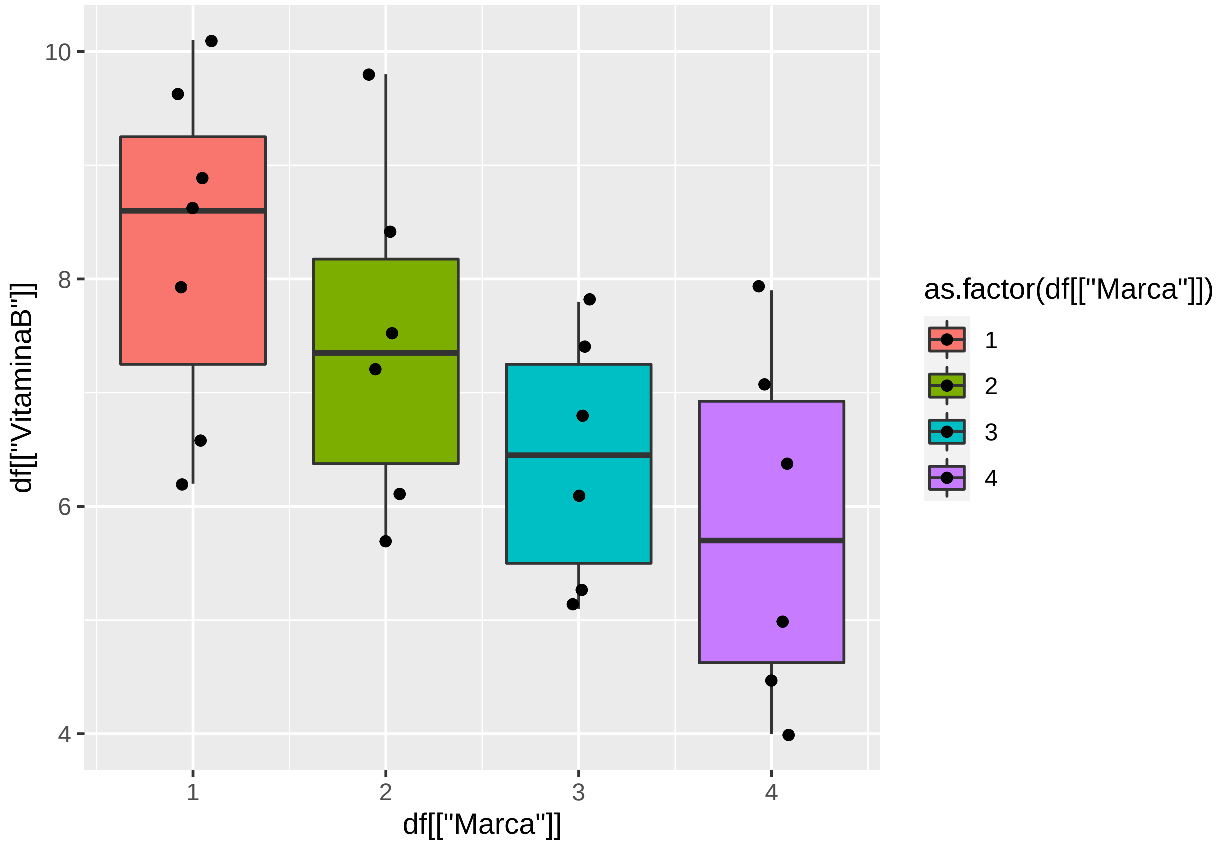
Head:

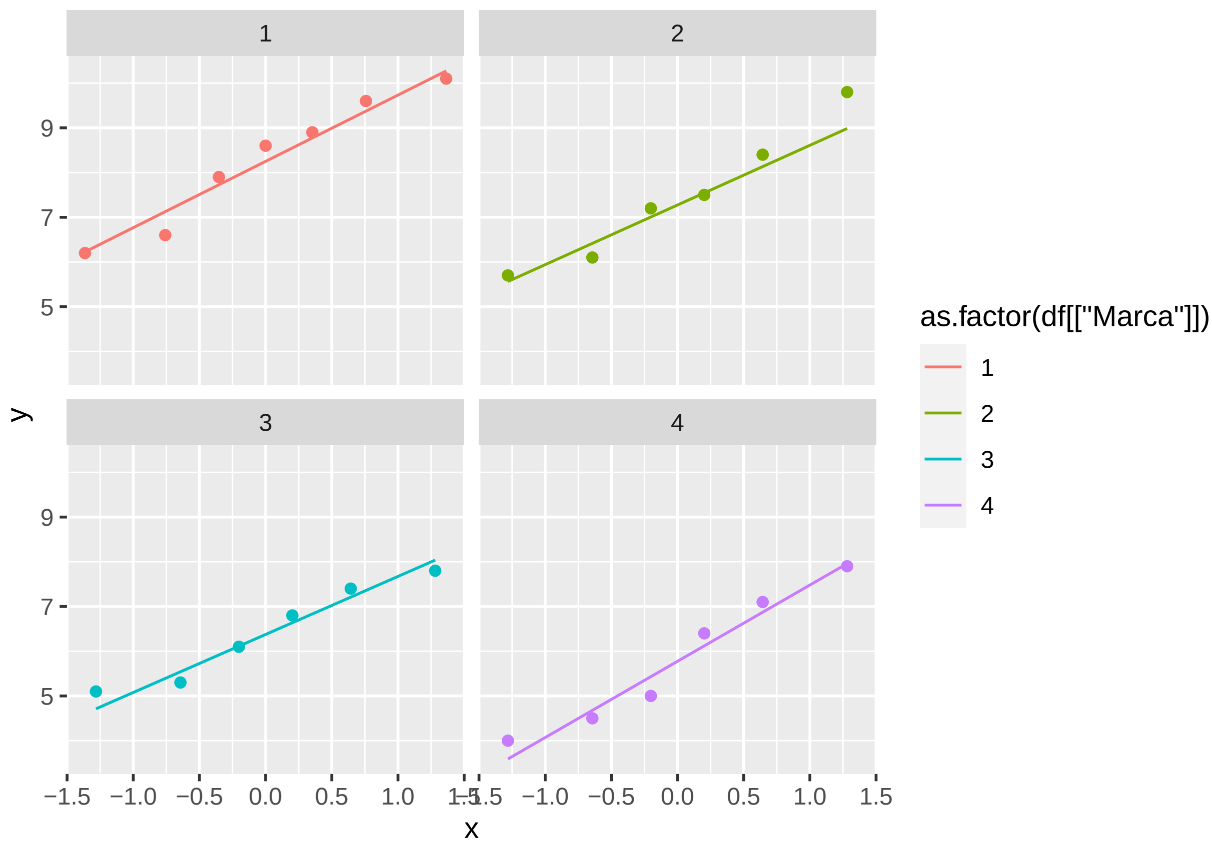
```
## # A tibble: 6 x 2
##   VitaminaB Marca
##       <dbl> <dbl>
## 1       7.9     1
## 2       6.2     1
## 3       6.6     1
## 4       8.6     1
## 5       8.9     1
## 6      10.1     1
```

Observaciones por grupo:

```
## # A tibble: 4 x 2
##   `df[["Marca"]]` `n()`
##             <dbl> <int>
## 1               1     7
## 2               2     6
## 3               3     6
## 4               4     6
```

Se cumplen los supuestos para su implementación?





Anova

```
##  [1]  7.9  6.2  6.6  8.6  8.9 10.1  9.6  5.7  7.5  9.8  6.1  8.4  7.2  6.8  7.8
## [16]  5.1  7.4  5.3  6.1  6.4  7.1  7.9  4.5  5.0  4.0
```

```
##  [1] 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4
## Levels: 1 2 3 4
```

fit del modelo

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fact         3  22.93   7.645   3.791 0.0256 *
## Residuals   21  42.35   2.016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

coeficientes

```
## (Intercept)       fact2       fact3       fact4
##   8.2714286  -0.8214286  -1.8547619  -2.4547619
```

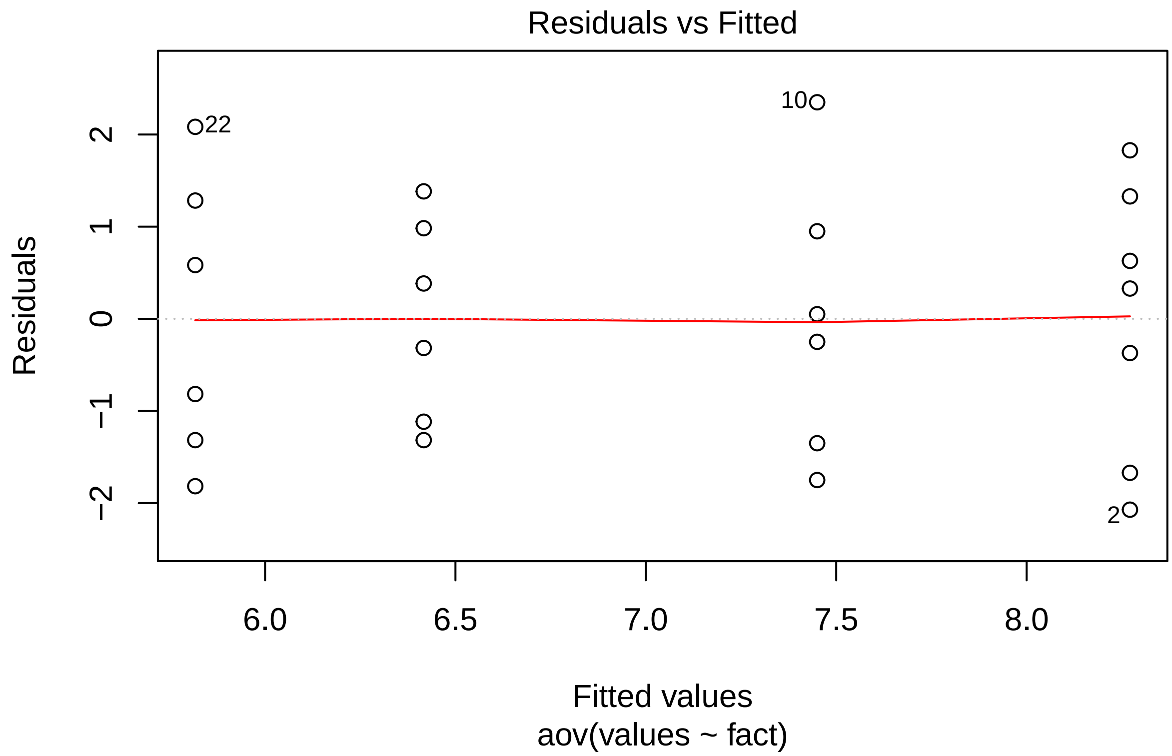
p-value

```
## [1] 0.02563266
```

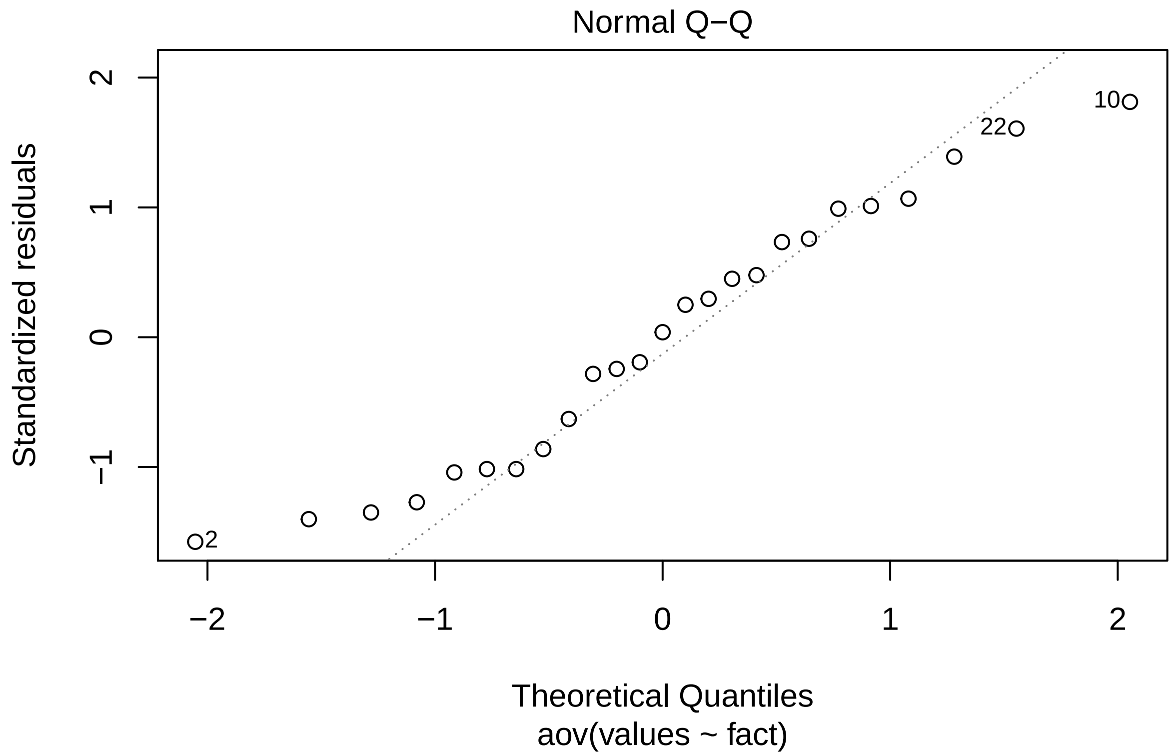
F-value

```
## [1] 3.791114
```

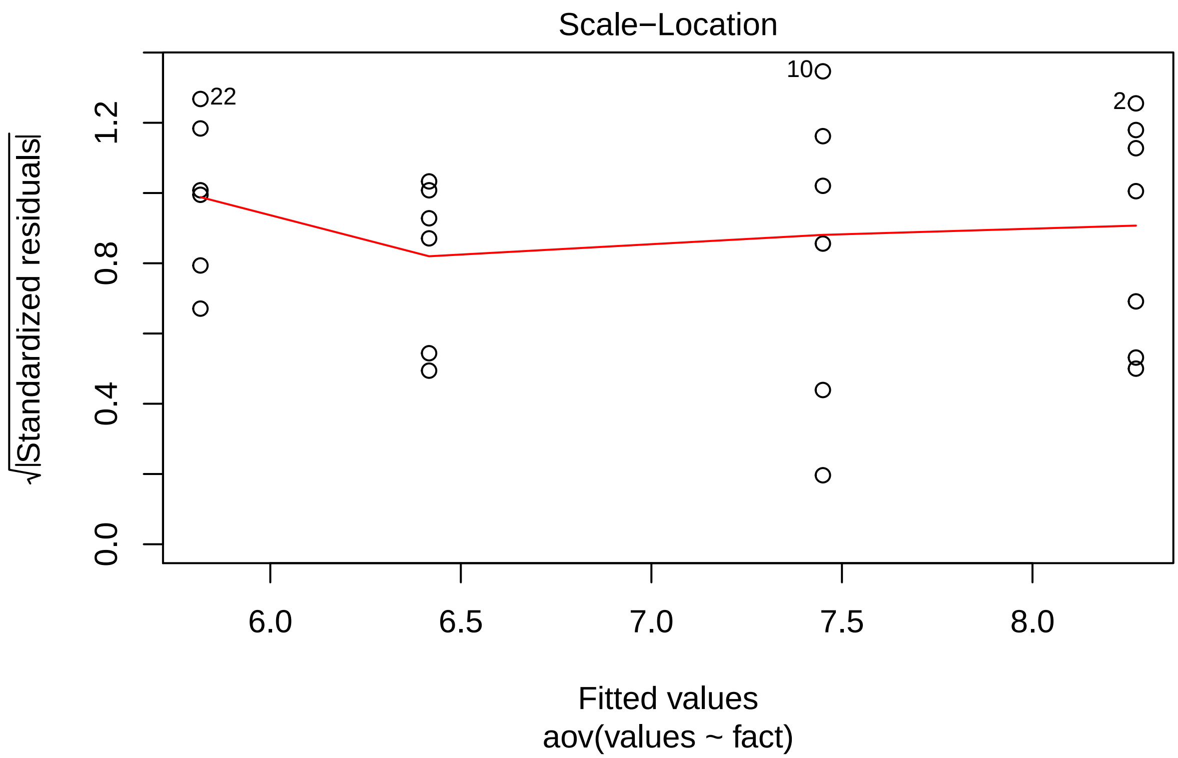
Plot ANOVA



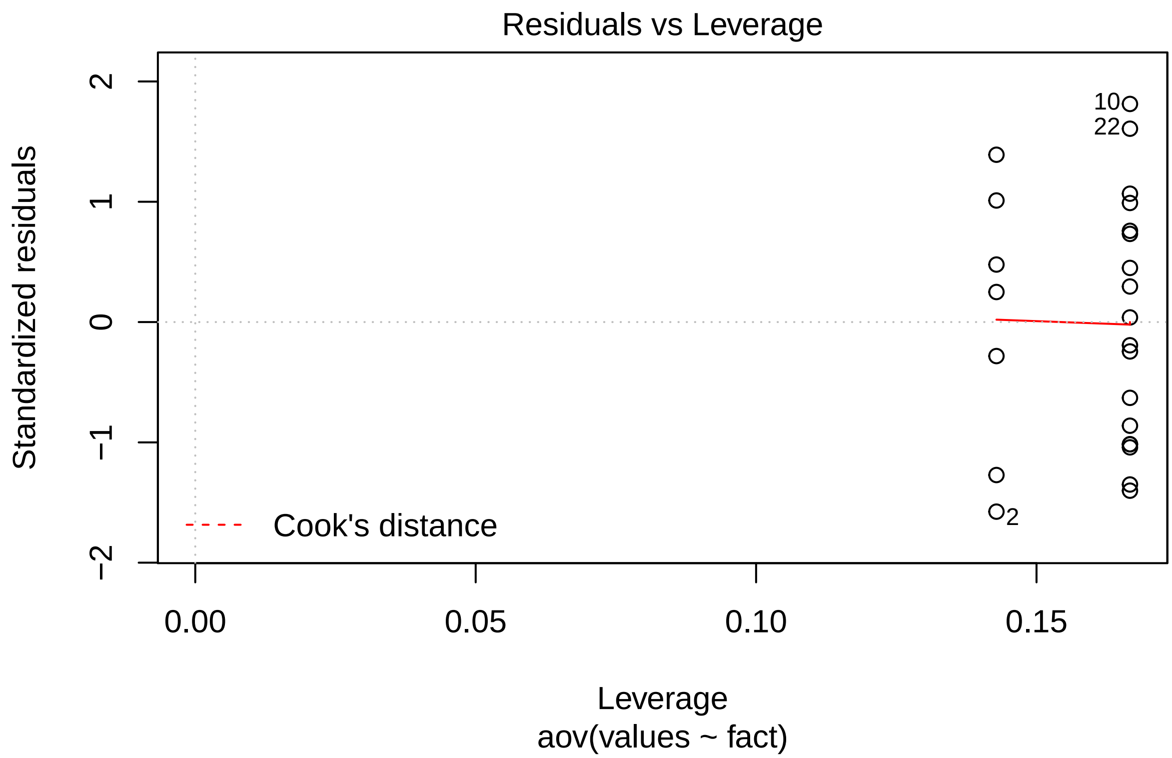
Residuals vs Fitted
Residuals
Fitted values
aov(values ~ fact)
22
10
2
6.0
6.5
7.0
7.5
8.0
−2
−1
0
1
2



Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(values ~ fact)
10
22
2
−2
−1
0
1
2



Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
22
10
2
6.0
6.5
7.0
7.5
8.0
Fitted values
aov(values ~ fact)



Conclusión

```
## [1] "H0 debe rechazarse, al menos dos medias son distintas a nivel de significancia 0.05"
```

Testear homosedasticidad

Test de Levene

que no es sensible a la falta de normalidad o a la presencia de valores atípicos

```
## Levene's Test for Homogeneity of Variance (center = "median")
##       Df F value Pr(>F)
## group  3  0.2949 0.8286
##       21
```

```
## [1] "No hay evidencia para rechazar H0, luego los datos son homosedásticos"
```

Test de Bartlett

sensibilidad al supuesto de normalidad

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df[[{     {         variable_numerica     } }]] by df[[{     {         variable_factor     }
## Bartlett's K-squared = 0.6168, df = 3, p-value = 0.8926
```

```
## [1] "No hay evidencia para rechazar H0, luego los datos son homosedásticos"
```

Testear normalidad

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(df_anova)
## W = 0.95307, p-value = 0.2937
```

```
## [1] "No hay evidencia para rechazar H0, luego los datos son normales"
```

Testear normalidad analizando residuos

```
##
##  Anderson-Darling normality test
##
## data:  residuals(df_anova)
## A = 0.36947, p-value = 0.3995
```

```
## [1] "No hay evidencia para rechazar H0, luego los datos son normales"
```

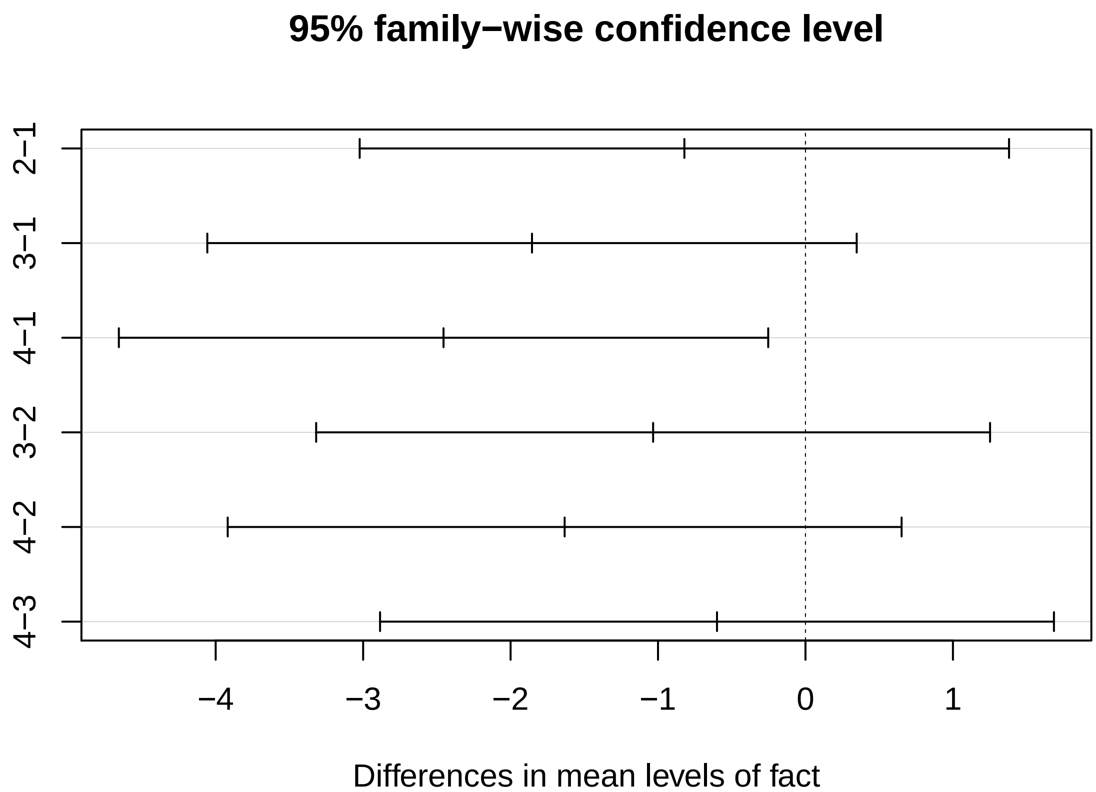
```
##
##  D'Agostino skewness test
##
## data:  residuals(df_anova)
## skew = 0.065564, z = 0.159996, p-value = 0.8729
## alternative hypothesis: data have a skewness
```

```
## [1] "No hay evidencia para rechazar H0, luego los datos son normales"
```

Anova y después: post-hoc

Tukey's Honest Significant Differences (HSD)

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ fact, data = df)
##
## $fact
##           diff       lwr        upr     p adj
## 2-1 -0.8214286 -3.023500  1.3806433 0.7284998
## 3-1 -1.8547619 -4.056834  0.3473100 0.1190380
## 4-1 -2.4547619 -4.656834 -0.2526900 0.0253506
## 3-2 -1.0333333 -3.318531  1.2518647 0.5969955
## 4-2 -1.6333333 -3.918531  0.6518647 0.2225040
## 4-3 -0.6000000 -2.885198  1.6851980 0.8832095
```



Cuando ANOVA no funciona: test de Kruskal-Wallis

```
##
##  Kruskal-Wallis rank sum test
##
## data:  values by fact
## Kruskal-Wallis chi-squared = 7.9977, df = 3, p-value = 0.04606

## [1] "H0 debe rechazarse: se encuentra significancia en la diferencia de al menos dos grupos"

## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##       obs.dif critical.dif difference
## 1-2  3.916667     10.80265      FALSE
## 1-3  8.416667     10.80265      FALSE
## 1-4 10.583333     10.80265      FALSE
## 2-3  4.500000     11.21044      FALSE
## 2-4  6.666667     11.21044      FALSE
## 3-4  2.166667     11.21044      FALSE
```

Analisis Discriminante Lineal (LDA)